Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials

ST John*,† and Gábor Csányi*,‡

†*Department of Physics, Cavendish Laboratory, University of Cambridge, Cambridge, CB3 0HE, United Kingdom and PROWLER.io, 66-68 Hills Road, Cambridge, CB2 1LA, United Kingdom*

‡*Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, United Kingdom*

E-mail: st@prowler.io; gc121@cam.ac.uk

Abstract

We introduce a computational framework that is able to describe general many-body coarse-grained (CG) interactions of molecules and use it to model the free energy surface of molecular liquids as a cluster expansion in terms of monomer, dimer, and trimer terms. The contributions to the free energy due to these terms are inferred from all-atom molecular dynamics (MD) data using Gaussian Approximation Potentials, a type of machine-learning model that employs Gaussian process regression. The resulting CG model is much more accurate than those possible using pair potentials. While slower than the latter, our model can still be faster than all-atom simulations for solvent-free CG models commonly used in biomolecular simulations.

1 Introduction

Simulations are a powerful tool to help us understand the behaviour of molecules. Molecular simulations with atomistic detail promise the highest accuracy, but are limited in the simulation length and time scales that can be reached in reasonable computing time. Molecular simulations can be applied to larger systems and over longer time scales when the system is described at a lower, "coarse-grained" (CG) resolution: groups of several atoms are combined into single CG sites. This has the additional advantage that it allows us to study interactions at their appropriate scale. For example, when we are interested in the orientational relationships of entire molecules, the thermal fluctuations of the atoms only constitute noise that obscures the properties we want to observe.

Even though coarse-graining intrinsically leads to a loss of detail, in the case of bottom-up CG models that have a direct mapping to an underlying all-atom model, it is in principle possible to create a *consistent* model that yields the same probability distributions for the CG coordinates as the all-atom model. In fact, by definition, the consistent CG potential is given by the potential of mean force (PMF), i.e. the free energy surface. This is, in general, a many-body potential. Different approaches have been proposed to derive a CG potential that approximates the potential of mean force, either from structure[1] or from forces of the underlying all-atom model, e.g., the multiscale

coarse-graining/force-matching, MSCG/FM.[2–5] In the limit of a complete basis set and infinite data, both approaches recover the many-body PMF.[6] It has been shown that both approaches are connected via the generalised Yvon–Born–Green integral equation theory.[7–9]

In practice, a complete basis is infeasible, and in many cases the interactions in CG models are based on pair potentials (at least for non-bonded interactions). These interactions cannot, in general, capture the free energy surface accurately, and hence practical CG models typically either reproduce certain structural distributions such as the Radial Distribution Function (RDF) *or* approximate the many-body mean forces, and are unable to reproduce both at once. The reproduction of cross-correlations beyond the one-dimensional Radial Distribution Functions is a strong indicator of a good CG model.[10]

There is active research into more complex interactions that result in more accurate CG models, such as pairwise interactions that depend on the local density[11–14] and three-body interactions between triplets of CG sites.[15–17] However, this has only been applied to one-site CG models that cannot describe the orientation of monomer units. Liu and Ichiye[18] use a many-body representation, but only for implementing a physically motivated baseline (electric dipole interactions), not to do a fit of the CG model in that larger-dimensional space.

In this paper we introduce an approach that allows for general many-body interactions in terms of a molecule-based cluster expansion, which permits a systematic convergence towards a consistent CG model which captures the potential of mean force more accurately, and thereby reproduces both many-body mean forces *and* structure, even with a limited basis set and limited data. We focus on the approximation of the many-body PMF at a given, fixed state point.

In Section 2 we give a brief overview of consistency in CG models, based on an explicit mapping from all-atom to coarse-grained model, and how this relates the CG interactions to the free energy surface/potential of mean force of the CG coordinates. In Section 3 we introduce our framework to infer many-body interactions from all-atom training data. We first discuss how we can describe the many-body PMF using local multi-body terms, based on "descriptors", which are essentially decompositions of the CG coordinates into local groups. In contrast to previous

work using a systematic many-body expansion of the PMF in terms of individual CG sites,[19,20] our approach is based on a cluster expansion in terms of entire molecules: each monomer can consist of several sites. Using such molecular many-body expansions is standard in chemistry, when fitting *potential energy surfaces*; indeed, we have made use of it previously when fitting models to quantum chemical data.[21] In Section 3.1 we define the monomer-based descriptors used throughout this work. In Section 3.2 we demonstrate the validity of approximating the free energy as a sum of local terms. These local contributions are multi-dimensional functions, and we use Gaussian process regression as a convenient way to infer them from noisy data (Section 3.3). Together with the descriptors, this constitutes the Gaussian Approximation Potential (GAP) approach. This framework was successfully applied before to infer all-atom potential energy functions (force fields) from corresponding quantum mechanical data.[22–26] In the case of applying GAP to coarse-graining (GAP-CG), the available samples of the true function behaviour are free energy gradients, i.e. derivatives of the PMF, in the form of instantaneous collective forces[27] or their average over the atomistic fluctuations, the mean forces. Learning a function only from very noisy observations of its derivatives leads to specific challenges to GAP-CG. We discuss insufficient sampling due to excluded volume in Section 3.4. In Section 3.5 we explain how we generate interaction potentials in GAP-CG and discuss computational details. We demonstrate the power of the GAP-CG approach in Section 4 on bulk methanol and bulk benzene, and on a solvent-free CG model of a benzene dimer in water. In Section 5 we conclude with a summary and an outlook to future work. A much more detailed description of this work is presented in a doctoral thesis.[28]

2 Consistent Coarse-Graining

Coarse-grained models are by construction approximations to the atomistically detailed behaviour of a system. We want to develop accurate, *consistent* coarse-grained models. As discussed by Voth et al.,[4] a coarse-grained model is consistent with an all-atom model when it gives probability distributions of the CG coordinates that correspond to those obtained with the all-atom model.

To be able to assess a coarse-grained model's accuracy and compare its predictions with an underlying all-atom model, we require a mapping function between the two that calculates the CG coordinates as a function of the atomic positions.

In the following, we consider an all-atom model with n atoms whose positions are denoted by the $3n$-dimensional vector $\boldsymbol{r}^n = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n\}$, and a coarse-grained model with N CG sites whose positions are denoted by the $3N$-dimensional vector $\boldsymbol{R}^N = \{\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N\}$. In general, upper-case variables will refer to the CG model and lower-case variables to the all-atom model.

The mapping is denoted by

$$\boldsymbol{R}^N = \boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n), \tag{1}$$

which describes the set of functions $\{\boldsymbol{M}_I\}$ that define CG coordinates in terms of the all-atom configuration, $\boldsymbol{R}_I = \boldsymbol{M}_I(\boldsymbol{r}^n)$, where $I = 1, \ldots, N$ indexes the CG coordinates. In principle, this may be an arbitrarily complex function. However, in practice this is usually a linear mapping, which can be used to describe, for example, centre of mass or centre of geometry of a group of atoms. The linear mapping is given by

$$\boldsymbol{M}_I = \sum_{i=1}^{n} c_{Ii} \boldsymbol{r}_i, \tag{2}$$

where i indexes the atoms in the all-atom model and c_{Ii} is the $N \times n$ matrix of mapping coefficients. For translational invariance we require $\sum_i c_{Ii} = 1$.

In the canonical (NVT) ensemble, the probability of occurrence for an all-atom configuration $\boldsymbol{r}^n$ is given by

$$p_r(\boldsymbol{r}^n) \propto \mathrm{e}^{-u(\boldsymbol{r}^n)/k_\mathrm{B}T}, \tag{3}$$

where $u(\boldsymbol{r}^n)$ is the potential energy of the all-atom system, T is the temperature, and k_B is the Boltzmann constant. Similarly, the probability distribution of a CG configuration $\boldsymbol{R}^N$ in the CG model is given by

$$P_R(\boldsymbol{R}^n) \propto \mathrm{e}^{-U(\boldsymbol{R}^N)/k_\mathrm{B}T}, \tag{4}$$

where U is now the potential used in the CG simulation.

However, we can also consider the probability distribution for finding a given CG configuration in the all-atom ensemble, as follows:

$$p_R(\boldsymbol{R}^N) = \int \mathrm{d}\boldsymbol{r}^n \, p_r(\boldsymbol{r}^n) \, \delta\big(\boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N\big), \tag{5}$$

where

$$\delta\big(\boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N\big) \equiv \prod_{I=1}^{N} \delta\big(\boldsymbol{M}_{\boldsymbol{R}I}(\boldsymbol{r}^n) - \boldsymbol{R}_I\big)$$

is the $3N$-dimensional Dirac delta function, ensuring that we include only those all-atom configurations that map to the given CG configuration. Equation (5) amounts to integrating out internal degrees of freedom.

For a consistent CG model, the probability for a CG configuration is the same in the all-atom and in the CG model:

$$P_R(\boldsymbol{R}^N) = p_R(\boldsymbol{R}^N). \tag{6}$$

This is equivalent to

$$\mathrm{e}^{-U(\boldsymbol{R}^N)/k_\mathrm{B}T} \propto \int \mathrm{d}\boldsymbol{r}^n \; \mathrm{e}^{-u(\boldsymbol{r}^n)/k_\mathrm{B}T} \; \delta\big(\boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N\big),$$

thereby defining the CG potential U. We can explicitly write U as a conditional free energy:

$$U(\boldsymbol{R}^N) = -k_\mathrm{B}T \log Z(\boldsymbol{R}^N) + (\text{const.}), \tag{7}$$

where

$$Z(\boldsymbol{R}^N) \equiv \int \mathrm{d}\boldsymbol{r}^n \; \mathrm{e}^{-u(\boldsymbol{r}^n)/k_\mathrm{B}T} \; \delta\big(\boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N\big) \tag{8}$$

is the constrained partition function as a function of the CG coordinates. Hence the CG interaction potential for a consistent CG model corresponds to the potential of mean force $A(\boldsymbol{R}^N) = -k_\mathrm{B}T \log Z(\boldsymbol{R}^N)$ for the $3N$ CG coordinates, viewed as collective variables.

The coarse-grained force on a CG site I is given by

$$\boldsymbol{F}_I(\boldsymbol{R}^N) = -\frac{\partial U(\boldsymbol{R}^N)}{\partial \boldsymbol{R}_I}. \tag{9}$$

This is equivalent to the mean force $-\partial A/\partial \boldsymbol{R}_I$ in the all-atom model. The mean forces can be directly measured through methods such as constrained MD[29,30] or restrained MD.[31,32] In this way we can determine the gradients of the potential of mean force to, in principle, arbitrary accuracy, and subsequently integrate the mean forces to obtain the CG potential.

It is also possible to determine free energy gradients directly from atomic forces in an unrestrained simulation. The "instantaneous collective force" (ICF),[27] $\boldsymbol{\mathcal{F}}_I$, is a noisy estimator of the free energy gradient with respect to the position of CG site I and is independent of any bias used in the simulation. In general, the ICF is a function of the atomic forces $\boldsymbol{f}_i = -\partial u(\boldsymbol{r}^n)/\partial \boldsymbol{r}_i$ for $i = 1, \ldots, n$ and, for arbitrary mapping functions, may explicitly depend on the configuration $\boldsymbol{r}^n$. For the linear mapping (2), the ICF is a linear combination of the atomic forces.[4] In the following, we will only consider centre-of-mass mappings from a group of atoms to a CG site, where each atom is part of at most one CG site. In this case, the ICF on a CG site I is simply the sum of the atomic forces on all atoms $i \in \mathcal{I}_I$ that are involved in the definition of that site ($c_{Ii} \neq 0$):

$$\boldsymbol{\mathcal{F}}_I(\boldsymbol{r}^n) = \sum_{i \in \mathcal{I}_I} \boldsymbol{f}_i. \tag{10}$$

Although the ICF is a deterministic function of the atomic coordinates $\boldsymbol{r}^n$, it can equivalently be viewed as a *noisy* sample of a stochastic function of just the CG coordinates, $\boldsymbol{\mathcal{F}}_I(\boldsymbol{R}^N)$. The ICF is the same as the "local mean force" used in the Adaptive Biasing Force method.[33]

The mean force (9) is the gradient of the free energy surface, and as such a function of the CG coordinates $\boldsymbol{R}^N$. It corresponds to the mean of the stochastic estimator $\boldsymbol{\mathcal{F}}_I(\boldsymbol{R}^N)$, and is given by the conditional canonical average over the ICF as a function of the atomic coordinates $\boldsymbol{\mathcal{F}}_I(\boldsymbol{r}^n)$,[4]

$$\boldsymbol{F}_I(\boldsymbol{R}^N) = \left\langle \boldsymbol{\mathcal{F}}_I(\boldsymbol{R}^N) \right\rangle = \left\langle \boldsymbol{\mathcal{F}}_I(\boldsymbol{r}^n) \right\rangle_{\boldsymbol{R}^N}, \tag{11}$$

where the subscript $\boldsymbol{R}^N$ indicates that only the all-atom configurations that map to this CG configuration are included in the average, weighted by the Boltzmann factor $\mathrm{e}^{-u(\boldsymbol{r}^n)/k_\mathrm{B}T}$:

$$\left\langle \bullet(\boldsymbol{r}^n) \right\rangle_{\boldsymbol{R}^N} \equiv \frac{\int \mathrm{d}\boldsymbol{r}^n \ \mathrm{e}^{-u(\boldsymbol{r}^n)/k_\mathrm{B}T} \ \delta\big(\boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N\big) \ \bullet(\boldsymbol{r}^n)}{\int \mathrm{d}\boldsymbol{r}^n \ \mathrm{e}^{-u(\boldsymbol{r}^n)/k_\mathrm{B}T} \ \delta\big(\boldsymbol{M}_{\boldsymbol{R}}^N(\boldsymbol{r}^n) - \boldsymbol{R}^N\big)}.$$

This converges to the same result as the explicit mean force calculation from a restrained or constrained simulation.[27]

Though Eqs. (7) and (8) formally define the CG potential, this is in terms of an average within the all-atom model. To be able to simulate the CG model on its own, without resorting to all-atom simulations in each step, we need to find a closed form for the CG potential, explicitly in terms of the CG coordinates. Hence, the practical challenge in coarse-graining is to determine an explicit function $U_\mathrm{CG}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N)$ of the CG coordinates that matches Eq. (7); generally, this will be an approximation to the true many-body PMF. This approximation depends on the functional form of U_CG and its parameters, and on the utility function that is used to optimise these parameters.

This utility function can be based on matching structural distributions of the all-atom model (Iterative Boltzmann Inversion,[34] Inverse Monte Carlo,[35–37] Relative Entropy[1]) or by matching the derivatives of the potential of mean force/the mean forces as in Eq. (11). We follow Voth et al. in aiming to match the free energy gradients. In principle, the MSCG/FM method can make use of arbitrarily complex basis functions, including functions of multiple scalar variables[6] and without having to define the interactions on fixed grids.[38] However, in practice, implementations of MSCG/FM rely on basis functions for distances between pairs (two-body interactions) and angles between triplets (three-body interactions) of CG sites. In contrast, in our framework we can tackle more complex interactions that involve entire molecules as a whole and describe two- and three-body interactions between *entire* molecules rather than only considering interactions on the level of individual CG sites. While this still constitutes a linear model, the basis functions in our framework are not predefined and can be more flexibly assigned depending on the underlying distribution of configurations.

3 GAP-CG

The parametrised CG potential that describes the free energy surface U is in principle a many-body potential of the $3N$ CG coordinates. However, determining U_{CG} as a general function of $3N$ arguments would be impractical. All coarse-graining approaches make an approximation to this; commonly, contributions to the CG potential are split into bonded and non-bonded terms,

$$U_{\text{CG}}(\boldsymbol{R}^N) = U_{\text{CG}}^{\text{bonded}} + U_{\text{CG}}^{\text{non-bonded}}. \tag{12}$$

Bonded terms include bond length, bond angle, and dihedral terms, and can be written as follows:

$$U_{\text{CG}}^{\text{bonded}} = \sum_{I,J}^{\text{bonds}} W^{\text{B}}(R_{IJ}) + \sum_{I,J,K}^{\text{angles}} W^{\text{A}}(\theta_{IJK}) + \sum_{I,J,K,L}^{\text{dihedrals}} W^{\text{D}}(\chi_{IJKL}), \tag{13}$$

where W^{B}, W^{A} and W^{D} are one-dimensional functions of a scalar argument, and $R_{IJ}(\boldsymbol{R}_I, \boldsymbol{R}_J) = |\boldsymbol{R}_J - \boldsymbol{R}_I|$ is the scalar distance between two CG sites, $\theta_{IJK} = \measuredangle(\boldsymbol{R}_I, \boldsymbol{R}_J, \boldsymbol{R}_K)$ is the angle between three CG sites, and χ_{IJKL} is the dihedral angle between four CG sites. In general, there may be different types of interactions with different functional forms depending on the types of CG sites that are involved. Though angle and dihedral terms are interactions between three and four sites, respectively, and hence describe "many-body" interactions, their contribution to the interaction energy is only determined by a single, scalar value.

Non-bonded terms are typically described by radial pairwise interactions:

$$U_{\text{CG}}^{\text{non-bonded}} = \sum_{I,J}^{\text{non-bonded pairs}} W^{\text{NB}}(R_{IJ}), \tag{14}$$

where we made explicit that even though such terms also depend on the distance between two CG sites, the functional relationship will be different. Note that in each case only a small number of CG sites are involved. Generally, there is an explicit cutoff for non-bonded contributions; only pairs of CG sites within a certain cutoff radius, $R_{IJ} < R_{\text{cut}}$, contribute.

The flexibility/parametrisation of the model is within the one-dimensional functions W. These functions are then optimised to approximate the PMF, $A \approx U_{\mathrm{CG}}$, based on one of a range of utility functions.

All these contributions can be written together as

$$U_{\mathrm{CG}} = \sum_{\xi} \sum_{M \in \mathcal{N}^{\xi}} W^{\xi}\big(D^{\xi}(\{\boldsymbol{R}_I | I \in \mathcal{N}_M^{\xi}\})\big), \tag{15}$$

where ξ iterates over the different types of interactions, the so-called "neighbourhoods" $\mathcal{N}_M^{\xi}$ are sets of CG sites that each contribute to a given interaction, and D^{ξ} is a scalar function of the positions of the CG sites within a given neighbourhood that gives their distance, angle, etc. The D^{ξ} characterise the argument to the contribution to the potential of mean force and we call them "descriptors", since together they characterise the neighbour environment of a CG site. For example, the descriptor for interactions between pairs of CG sites is the radial distance, $D^{\mathrm{B}}(\boldsymbol{R}_I, \boldsymbol{R}_J) = |\boldsymbol{R}_J - \boldsymbol{R}_I|$, and the descriptor for the bonded angle interaction is the angle between three sites, $D^{\mathrm{A}}(\boldsymbol{R}_I, \boldsymbol{R}_J, \boldsymbol{R}_K) = \measuredangle(\boldsymbol{R}_I, \boldsymbol{R}_J, \boldsymbol{R}_K)$.

We now extend this to vector-valued descriptors $\boldsymbol{D}^{\xi}$; this means that in Eq. (15) we replace the scalar functions D^{ξ} with vector-valued functions $\boldsymbol{D}^{\xi}$, so that the local free energy contributions W^{ξ} can now be functions of more than one argument.

3.1 Descriptors

In this work, our descriptors are based on a cluster expansion of monomers. Analogous to a cluster expansion of the potential energy, we can approximate the PMF by a molecular many-body cluster expansion using a sum of monomer-based terms:

$$\begin{aligned} A(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N) = A(\boldsymbol{R}_{\{i,\alpha\}}) \approx U_{\mathrm{CG}} = \sum_j W_{\text{mono}}\big(\boldsymbol{D}_{\text{mono}}(\boldsymbol{R}_{j,\{\alpha\}})\big) \\ + \sum_{j,k} W_{\text{dimer}}\big(\boldsymbol{D}_{\text{dimer}}(\boldsymbol{R}_{j,\{\alpha\}}, \boldsymbol{R}_{k,\{\beta\}})\big) + \sum_{j,k,l} W_{\text{trimer}}\big(\boldsymbol{D}_{\text{trimer}}(\boldsymbol{R}_{j,\{\alpha\}}, \boldsymbol{R}_{k,\{\beta\}}, \boldsymbol{R}_{l,\{\gamma\}})\big), \end{aligned} \tag{16}$$

where Roman letters i, j, k, l iterate over molecules and Greek letters α, β, γ iterate over CG sites within each molecule. The monomer, dimer, and trimer descriptors are constructed from *all* pairwise distances between the involved CG sites. The monomer, for example, is defined by the set of all intramolecular distances:

$$\boldsymbol{D}_{\text{mono}}(\boldsymbol{R}_{j,\{\alpha\}}) = \{D^{jj}_{\alpha\beta} \mid \alpha = 1, \ldots, M;\ \beta = \alpha + 1, \ldots, M\} \tag{17}$$

where $D^{jk}_{\alpha\beta} = |\boldsymbol{R}_{j,\alpha} - \boldsymbol{R}_{k,\beta}|$ is the distance between site α in molecule j and site β in molecule k, and M is the number of CG sites within a monomer. For a two-site monomer, this is equivalent to the one-dimensional potential describing the CG bond. For a three-site monomer, we have $\boldsymbol{D}^{j}_{\text{mono}} = \{D^{jj}_{12}, D^{jj}_{13}, D^{jj}_{23}\}$. This means the intramolecular CG potential is a function of three distances, $W_{\text{mono}}(D_{12}, D_{13}, D_{23})$. Note that this corresponds to a much larger, more flexible class of functions than a typical pair potential, which is a sum of three functions depending on the distances individually, $W_{12}(D_{12}) + W_{13}(D_{13}) + W_{23}(D_{23})$. In the approaches found in the literature, the energy of a three-site monomer, for example, would commonly be written as a sum of terms that depend on the two bond lengths and their angle:

$$W^{\text{B}}_1\big(D^{\text{B}}(\boldsymbol{R}_I, \boldsymbol{R}_J)\big) + W^{\text{B}}_2\big(D^{\text{B}}(\boldsymbol{R}_J, \boldsymbol{R}_K)\big) + W^{\text{A}}\big(D^{\text{A}}(\boldsymbol{R}_I, \boldsymbol{R}_J, \boldsymbol{R}_K)\big).$$

which, due to the particular way the three descriptors are separated, is again capable of representing only a subset of the possible functions of three variables.

The molecular dimer descriptor is defined by the ordered set of all intra- as well as intermolecular distances:

$$\boldsymbol{D}_{\text{dimer}}(\boldsymbol{R}_{j,\{\alpha\}}, \boldsymbol{R}_{k,\{\beta\}}) = \{D^{jj}\}, \{D^{jk}\}, \{D^{kk}\}. \tag{18}$$

For a two-site monomer, this consists of the following list of six distances:

$$(D^{jj}_{12}, D^{jk}_{11}, D^{jk}_{12}, D^{jk}_{21}, D^{jk}_{22}, D^{kk}_{12}),$$

or, written out:

$$(|\boldsymbol{R}_1^j - \boldsymbol{R}_2^j|, |\boldsymbol{R}_1^j - \boldsymbol{R}_1^k|, |\boldsymbol{R}_1^j - \boldsymbol{R}_2^k|, |\boldsymbol{R}_2^j - \boldsymbol{R}_1^k|, |\boldsymbol{R}_2^j - \boldsymbol{R}_2^k|, |\boldsymbol{R}_1^k - \boldsymbol{R}_2^k|).$$

For a three-site monomer, the dimer descriptor contains 15 distances: the three intramolecular distances of the first monomer, the nine intermolecular distances, and the three intramolecular distances of the second monomer. The intramolecular distances within a dimer descriptor do *not* play the same role as the monomer term; they are needed to describe the deformation of molecules due to the interaction with another molecule.

Analogously, the trimer descriptor is defined by the set of all intra- as well as intermolecular distances of involved CG sites. For a single-site description of the CG molecules this contains three distances, for a two-site description this contains 15 distances, and for a three-site description this contains 36 distances.

When we approximate the free energy surface by a sum over monomer, dimer, and trimer terms, the contribution due to each term should only depend on the relative positions and orientations of the CG sites within that molecule, pair, or triplet — not on their absolute position or orientation with respect to a global coordinate system. This is already achieved by these distance-based descriptors. The potential of mean force should also be invariant under permutations of equivalent CG sites. For a traditional CG potential this is trivial, because the sum of terms in Eq. (13) is already permutationally invariant. However, the ordered list of site–site distances in our monomer-based descriptors is not yet permutationally invariant; in our framework we take permutational symmetry into account by explicitly summing the $W(\boldsymbol{D})$ functions over the permutation group. This is explained in more detail in Appendix A.1.

3.2 Free energy is local

The free energy surface is a function of all $3N$ CG coordinates, and describing it by a sum of local contributions constitutes an approximation. However, to a large extent the free energy is

observed to be a *local* function: similar to how the force on an atom only depends on the relative positions of nearby atoms, not on the positions of atoms that are far away, the mean force on a CG site, the derivative of the free energy, only depends on the positions of nearby CG sites. Even electrostatic interactions are often localised by screening. Coarse-grained models generally rely on the locality approximation, though this approximation may be made implicitly. Note that the locality of the many-body PMF is not given by the locality of the underlying interactions; in fact, cutoffs in CG simulations typically need to be longer than in atomistic models.

We can explicitly test the validity of this approximation in a "locality test". For this, we constrain the positions of a central CG site and all CG sites within a certain neighbourhood (defined by a cutoff radius). We then calculate the mean force on the central site for different configurations of the CG sites beyond the neighbourhood. Different beyond-neighbourhood configurations lead to different values of the mean force on the central site, but the locality of the free energy can be seen in that this variation becomes smaller and smaller as the cutoff is increased and the neighbourhood includes more and more CG sites.

This is demonstrated using the example of a two-site CG model for bulk methanol for two different cutoff radii in Fig. 1. To take into account statistical uncertainty, we repeat the procedure outlined in the following for 10 uncorrelated configurations sampled from an NVT MD simulation of the all-atom model (groups 1 to 10 in Fig. 1).

First, we select a snapshot from the all-atom simulation, and determine the corresponding CG configuration. We pick one molecule (the "central molecule") and constrain its CG coordinates and those of its immediate neighbourhood, defined as all molecules within a certain cutoff. Here we compare two cutoffs based on the their centres of mass (COM), 6.0 Å and 9.2 Å, corresponding to the first and second minima of the COM RDF, respectively. The smaller cutoff includes 12 to 17 molecules (average 14.0); the larger cutoff includes 44 to 55 molecules (average 50.5).

Second, we randomise the configuration of molecules outside the neighbourhood. To achieve this we run an all-atom simulation with the CG sites of the central molecule and all molecules in its neighbourhood fixed. The remaining molecules in the simulation box move freely, and from

each of these simulations we sample 16 configurations that correspond to identical CG coordinates for the central molecule and its neighbourhood but different coordinates for the other molecules.

Finally, to determine how the mean force on the central molecule depends on the changes of the environment outside the neighbourhood, for all 16 configurations, we carry out all-atom MD simulations in which all CG coordinates are constrained.

Overall, this leads to $10 \times 16 \times 2 = 320$ individual mean force evaluations, which are visualised in Fig. 1. This shows that the variation is significantly smaller for the larger cutoff radius, thereby explicitly motivating our local approximation. In the following, we assume that the locality approximation is valid in our test systems. An explicit locality test may be warranted when a CG model is unable to reproduce the mean forces of the underlying all-atom model to our desired accuracy. Note that for a given cutoff we cannot expect our CG model to approximate the mean forces any better than the inherent error due to the locality approximation that we found in the locality test.

3.3 Regression

In our GAP approach, we infer the multi-dimensional functions W using Gaussian process regression (this is reviewed in Appendix A). This allows us to determine the local contributions, even though in standard MD simulations we cannot observe them directly, only the (potentially very noisy!) *gradients* of the total *sum* of the local contributions. Unlike standard linear regression such as the least-squares fit in MSCG/FM, in Gaussian process regression the basis functions are not directly chosen; instead, statements and assumptions are made about the covariance structure of the unknown function, which follow easily from chemical intuition and simple tests (e.g., locality, as discussed above), and these together with the data distribution and type (function values, derivatives, sums, etc.) imply the set of basis functions, enabling the use of more complex, flexible representations. Moreover, Gaussian process regression is inherently *regularised* and hence results in smooth potentials even with sparse data. (In contrast to the Bayesian statistics approach to MSCG/FM,[39] the regularisation hyperparameter in Gaussian process regression can be interpreted

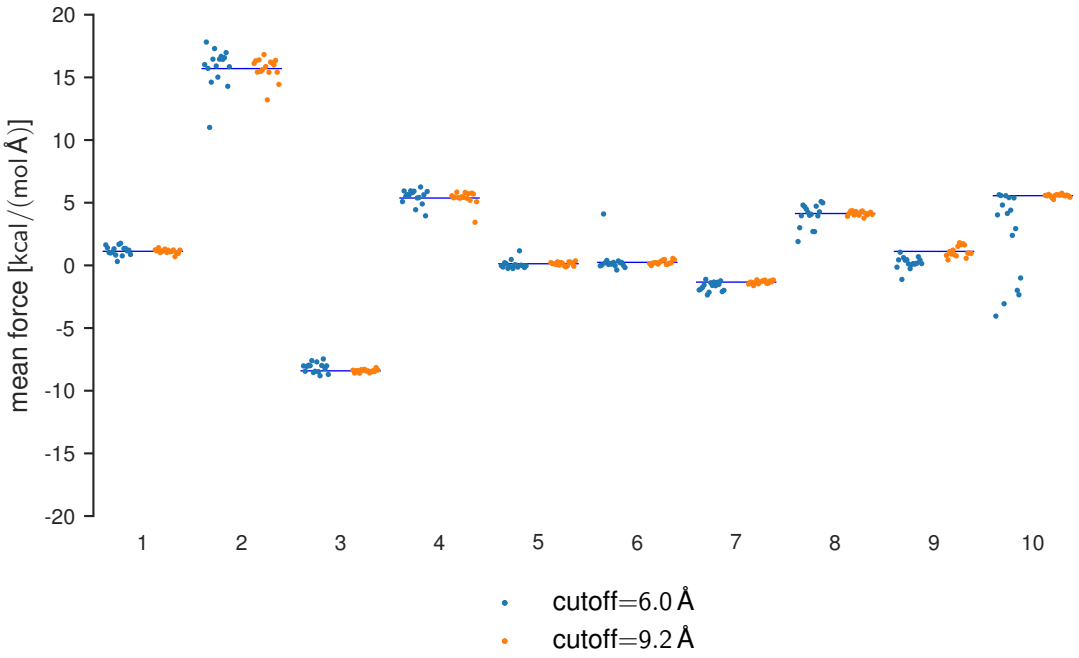

Figure 1: Locality test of the mean forces for methanol. Distribution of the mean force components, using the example of the z component of the mean force on the first CG site of the central molecule for 16 different configurations beyond the neighbour shell, for two different sizes of the neighbour shell (cutoffs of 6.0 Å and 9.2 Å, corresponding to the first and second minima of the COM RDF). Each group 1 to 10 corresponds to a different configuration of the inner environment. Each dot corresponds to a single mean force evaluation, depending on the configuration of the neighbourhood (the different groups) and the configuration of the environment outside the neighbourhood (the cluster of dots in each group). The horizontal lines denote the average over the mean force values for the larger cutoff in each group. See the main text for further details.

as a quantification of the noise in the data (both coming from statistical fluctuations and locality error) and has physical units, as opposed to an abstract quantity that needs to be optimised.)

The overall workflow of the GAP-CG approach is illustrated in Fig. 2. Analogously to interatomic potentials, the application of Gaussian processes to CG potentials significantly benefits from taking into account the underlying physics. Interactions do not have a uniform length scale, and how to address inhomogeneous length scales in Gaussian process regression is discussed in the Appendix (Appendix A.3). Moreover, the free energy surface will be undersampled at unfavourable configurations due to excluded volume; this will be discussed in the following section. In Section 3.5 we discuss setting the hyperparameters for the Gaussian process regression.

3.4 Excluded volume

In general, it is not possible to exhaustively enumerate the high-dimensional configuration space of a system using observations on grid points. Instead, we will need to sample training configurations from the underlying all-atom simulation. This works fine to allow us to infer the PMF around the (highly populated) free energy minima. However, regions of configuration space with high free energy will be undersampled. This is typically the case for configurations containing molecules at close range, which have a high potential energy, leading to a very small Boltzmann factor that describes a very small probability of finding the system in this state. The corresponding coarse-grained configurations should likewise have a small probability, and hence they are described by a high free energy.

The Gaussian process regression, away from training data, will regress to the prior mean function, which is zero, whereas the function should in fact go to large values. To deal with this, we include a repulsive pair potential with a short-range cutoff ("short-range repulsive pair potential", SRRPP) as part of the model. However, we do not have to fit this from the data and can determine it in some other way. This provides a baseline that can describe excluded volume due to the short-range repulsion of atoms and molecules. Useful baseline potentials may be based on the

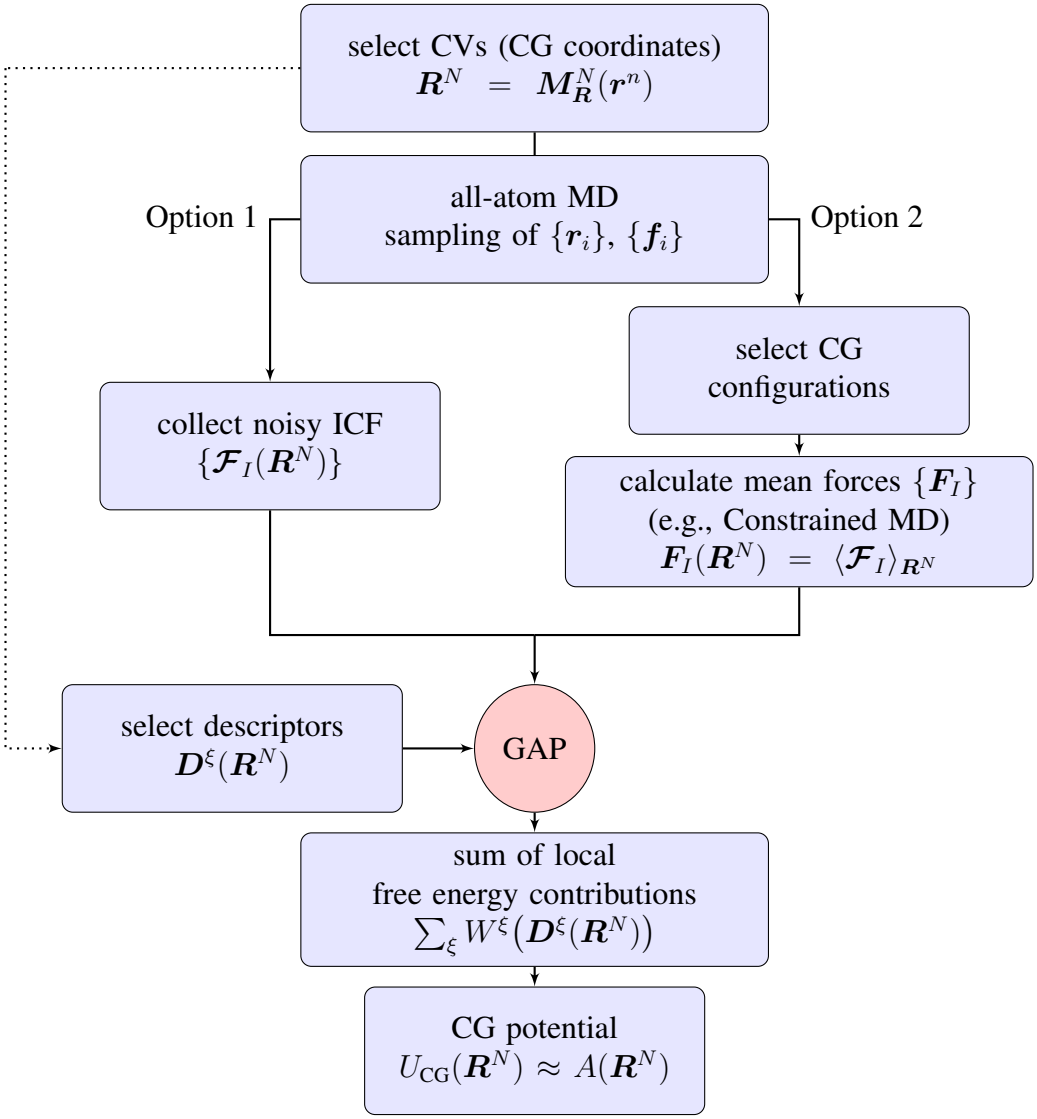

Figure 2: Workflow for GAP-CG

centre-of-mass distance between molecules, or descriptors for pairwise site–site distances. As for reference potentials used in other approaches,[40] the baseline potential is subtracted before training the GAP (by evaluating it for all training configurations and subtracting the corresponding forces from the training forces), and added again for prediction.

The baseline potential has two separate functions. First, it is a good approximation to the PMF at close approach. This means that the "difference surface" that needs to be described by the more complex descriptors is flattened out, which makes it easier to represent. Second, the baseline potential can ensure that the predicted free energy keeps increasing as CG sites are brought close, which guarantees that the CG model will not visit such unphysical configurations.

3.5 Hyperparameters

Developing a GAP model requires setting the hyperparameters, e.g., length scale of basis functions, regularisation parameters, etc. In principle, these could be optimised from the data itself;[41] however, for the large data sets required in our use case, this would be computationally expensive, and is not actually necessary: we can derive reasonable settings from physical principles and intuition. The values we use are given in Table S2 for methanol and in Table S3 for benzene. With a larger amount of training data, the influence of the prior, defined by the values of the hyperparameters, decreases. This is shown in Fig. 3 using the example of the assumed noise level of the training data.

The pair potentials are quickly saturated by data: their prediction error does not change significantly either with the noise level hyperparameter or the amount and noisiness of the training data. The dimer descriptor has much more expressive power, and can describe the true forces more accurately. This demonstrates that, with sufficient training data, training from noisy ICFs converges to the training from the already averaged free energy gradients (mean forces).

Importantly, this figure suggests choosing a very high noise level for a small and noisy training set. With an increasing number of training points, the error depends less and less on the precise value of the noise level. This implies that the minimum is not stable and, hence, it does not make

sense to "optimise" the noise level parameter; this would risk overfitting of the training data. As the prediction performance is stable with respect to the noise level for larger training sets, we can simply choose a sufficiently high noise level.

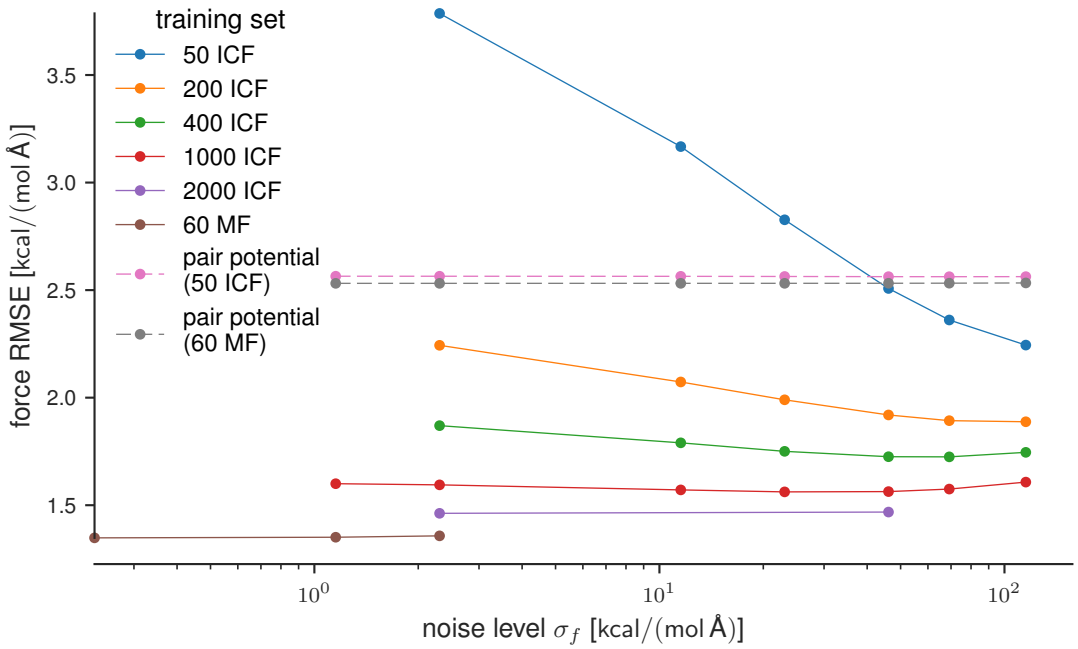

Figure 3: Convergence of force prediction error with increasing amount of data, at the example of three-site coarse-grained bulk benzene. We show the Root Mean Squared Error (RMSE) of the force prediction of GAP-CG models for different values of the noise level hyperparameter σ_f (setting the amount of regularisation); the "ground truth" is given by the actual mean forces, obtained through constrained MD in the all-atom model, on a separate test set (not used in the training of the potentials). The solid lines are for a GAP-CG model based on the 15-dimensional molecular dimer descriptor. The dashed lines are for a GAP-CG model using one-dimensional pair potentials, comparable to typical MSCG/FM models. (When fitting site–site pair potentials, these methods can represent broadly the same class of functions; the key distinction is the regularisation, which becomes less and less important in the limit of infinite data.) The different lines show the influence of the amount and type of training data: the noisy instantaneous collective forces (ICF), for 50 to 2000 many-body configurations of the all-atom model and the converged mean forces (MF) for 60 many-body configurations.

4 Results and Discussion

We apply the GAP-CG approach to model systems of methanol, described by a two-site model (with one site on the COM of the CH_3 group and one on the COM of the OH group), and of benzene, described by a three-site model (with one site on the centre of mass of each pair of neighbouring carbon atoms). Methanol and benzene are considered in their bulk state. For benzene, we further show results of a benzene dimer solvated by water.

We compare our GAP approach with commonly used state-of-the-art approaches using pair potentials. We include pair potentials derived from structure matching to Radial Distribution Functions (RDFs) and from matching forces. For structure matching we use the Iterative Boltzmann Inversion (IBI) procedure[34] as implemented in the VOTCA package.[42] For force-matching we use the MSCG/FM package provided by the Voth group.[4,5,43]

All-atom models were simulated using AMBER.[44] Pair potential CG models were simulated using LAMMPS.[45] GAP-CG simulations were run with QUIP[1]. Timings for the GAP interactions are given in Table 1.

Table 1: Computational cost of GAP many-body interactions for the two-site methanol and three-site benzene CG models discussed in the following, for the force evaluation of a single interaction on a single CPU core. The hyperparameters are given in Table S2 and Table S3. Each bulk system contains 216 molecules. The benzene interactions are more expensive because of the high symmetry of the molecule. The methanol monomer term is omitted, being equivalent to a standard pair potential.

system	interaction	dimensionality	time / ms	number of interactions
methanol	dimer	6	1.017(2)	1416(14)
	trimer	12	3.82(7)	500(27)
benzene	monomer	3	0.106(9)	216
	dimer	12	23.5(3)	1429(12)

[1]The GAP code is available for non-commercial use from `https://github.com/libatoms/quip/`.

4.1 Bulk methanol

We considered 216 methanol molecules in a periodic cubic box of side length $L = 24.205\,\text{Å}$, determined using NPT equilibration [cf. supporting information, Table S1]. For the force-based CG approaches (MSCG/FM and GAP), the training set was constructed as follows: 30 uncorrelated configurations of the atomistic model were sampled from unconstrained MD in the NVT ensemble in the condensed phase. For each of these all-atom configurations, we constrained the CG coordinates of all molecules using PMFlib[46] and collected the many-body mean forces in constrained MD runs of $100\,\text{ps}$ each. This resulted in mean force observations that were converged to within $0.1\,\text{kcal}/(\text{mol}\,\text{Å})$.

We applied our GAP framework using the combination of monomer + dimer and monomer + dimer + trimer descriptors. The corresponding hyperparameters are given in Table S2. For the combination of monomer + dimer descriptors, we used as a baseline a repulsive pair potential between the COM of molecules (cutoff=$3.45\,\text{Å}$), trained on COM mean force data for two molecules in vacuum. For the combination of monomer + dimer + trimer descriptors, the baseline repulsive pair potential was based on mean forces on CG sites as two molecules are brought together for the three combinations of CG sites (cutoff=$6\,\text{Å}$).

The MSCG/FM pair potentials showed no significant difference between using $6\,\text{Å}$ or $9\,\text{Å}$ cutoffs. Our results are equivalent to those in the paper by Izvekov and Voth.[3]

The structure-matching pair potentials were obtained by IBI as follows. The CG bond potential was obtained by Direct Boltzmann Inversion using the harmonic approximation. The non-bonded pair potentials were refined iteratively using the VOTCA package[42] (starting from Direct Boltzmann Inversion). The three pair potentials were defined on a grid up to a cutoff of $10\,\text{Å}$ with spacing $0.1\,\text{Å}$. To improve the stability of the IBI algorithm, potential updates were smoothed twice and scaled by a factor of 0.2. In each step, only one of the three non-bonded potentials was updated.

To show that a large contribution to the overall structure is simply due to geometric constraints, we include in the results a short-range repulsive pair potential (SRRPP) that is the short-range

part of the Boltzmann-inverted potential[47] from the pair distance distributions; the cutoffs were as follows: CH_3–OH: 3.6 Å; CH_3–CH_3: 4.0 Å; OH–OH: 2.85 Å.

While the structure-matching IBI pair potential perfectly reproduces, by construction, the site–site distance distribution, this does not extend to other distributions of the system, already including the centre-of-mass distance distribution, see Fig. 4. The pair potential based on matching forces does not provide a good reproduction of the RDFs. In contrast, the GAP-CG model, also based on matching forces, reproduces both site–site and COM RDF much more closely. Note that the long-range part of the structure is already described by the short-range repulsive pair potential.

The shortcomings of pair potentials become even more apparent when considering more complex distributions such as the Angular Distribution Function of OH sites (Fig. 5) or the orientational distribution of pairs of methanol molecules as a function of their COM distance (Fig. 6). In each case, the distributions in simulations with GAP can be seen to reproduce the all-atom reference distributions not just better than the pair potential based on matching forces, but also better than the pair potential optimised for matching structures.

4.2 Bulk benzene

We considered 216 benzene molecules in a periodic cubic box. Following equilibration [cf. supporting information, Table S1], the box length was determined to be $L = 32.0205$ Å.

For benzene, the combination of monomer + dimer descriptors proved to be sufficient to accurately capture the interactions. The hyperparameters are given in Table S3. The short-range repulsive pair potential (SRRPP) was obtained by fitting a pair potential with 4 Å cutoff to the mean force training data to capture the appropriate short-range repulsion, shown in Fig. S2. This was used as the baseline for the GAP training. IBI and MSCG/FM pair potentials were obtained as for methanol.

In Fig. 7 we show the orientational distribution of pairs of benzene molecules as a function of their COM distance. Uniformly random orientations would lead to a uniform profile of the dot product, corresponding to all lines in the right-hand plot lying on top of each other. The figure

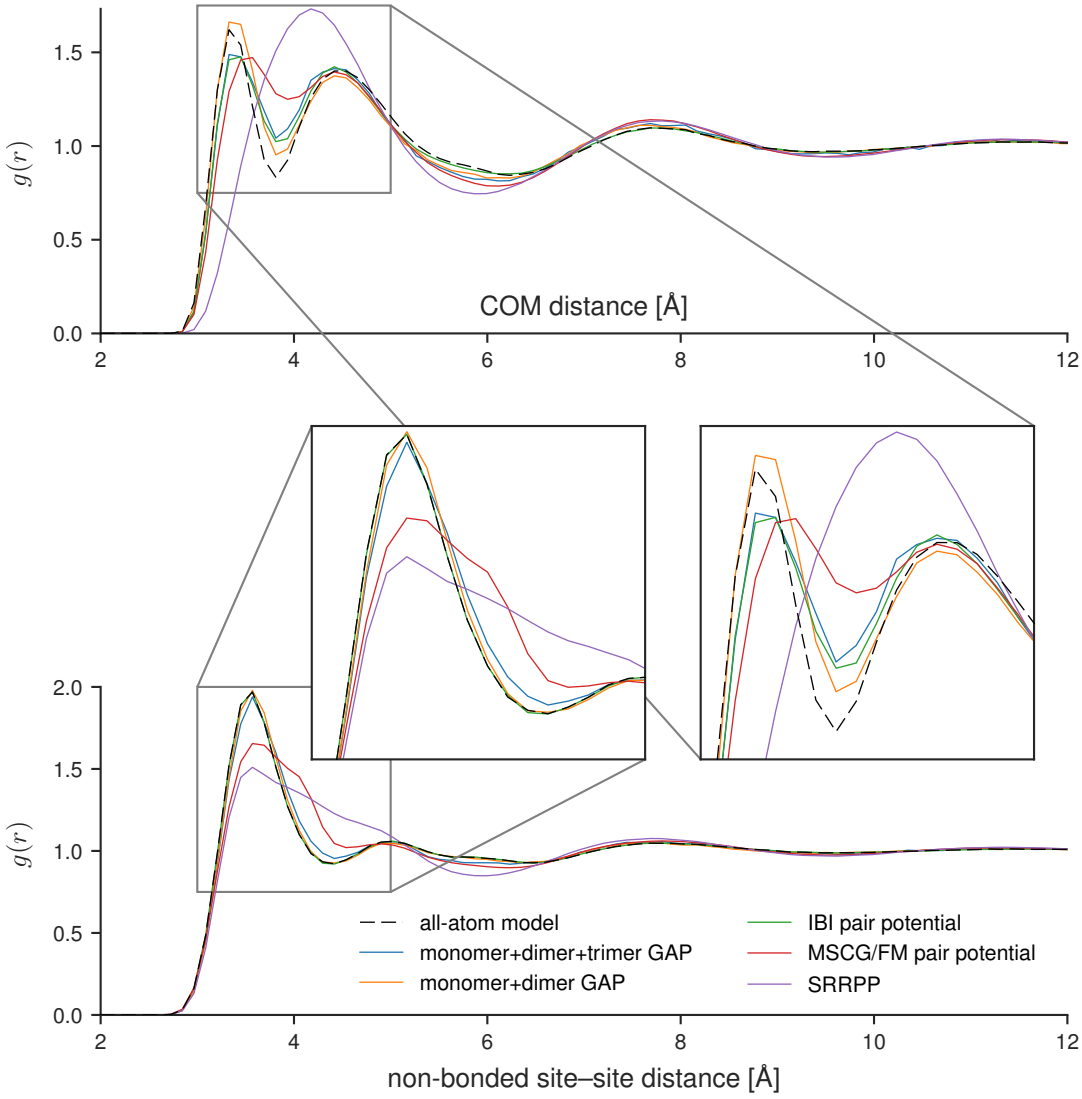

Figure 4: Radial Distribution Functions (RDFs) in bulk methanol, comparing the all-atom model and different two-site coarse-grained potentials. We compare the results using our GAP dimer and trimer descriptors with pair potentials based on matching structure (Iterative Boltzmann Inversion; IBI), matching mean forces (MSCG/FM), and short-range repulsive pair potentials (SRRPP) that describe excluded volume only. *Top:* centre-of-mass (COM) RDF. *Bottom:* RDF between CH_3 and OH site.

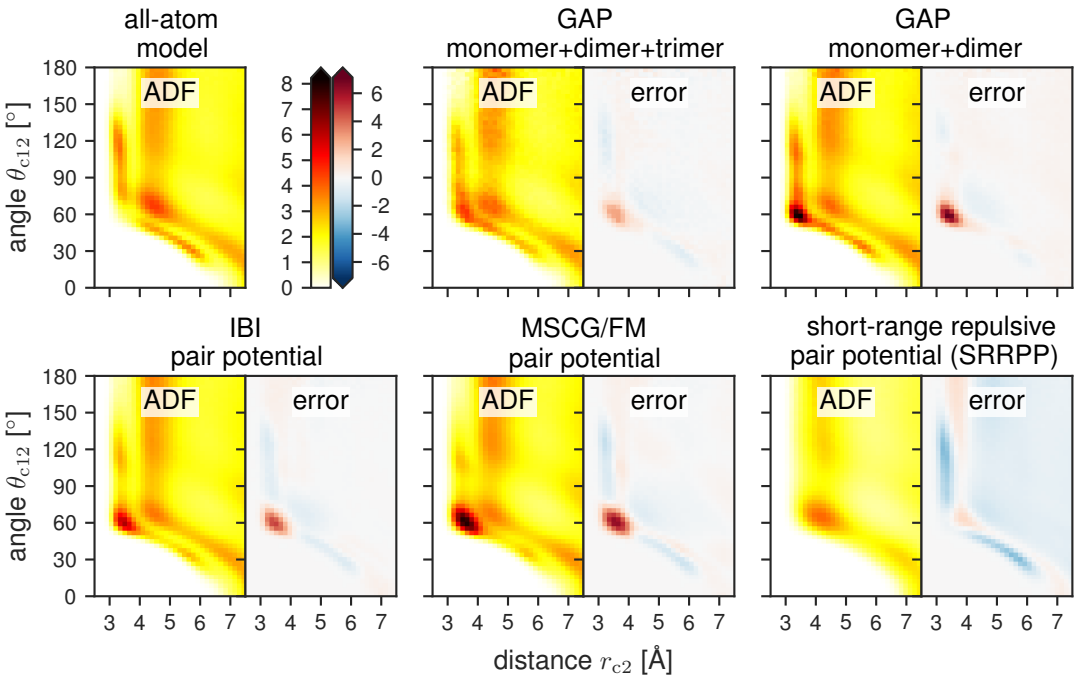

Figure 5: Angular Distribution Function (ADF) of molecule COMs in bulk methanol for different potentials. The ADF (left-hand plot of each pair) is determined by considering all triplets of molecules where the COM distance r_{c1} between a central molecule and its first neighbour is less than $3.8\,\text{Å}$, which indicates a hydrogen bond. Here we show the 2D distribution of the angle θ_{c12} at the central molecule against the COM distance r_{c2} between the central molecule and its second neighbour. The right-hand plot of each pair shows the deviation from the all-atom result.

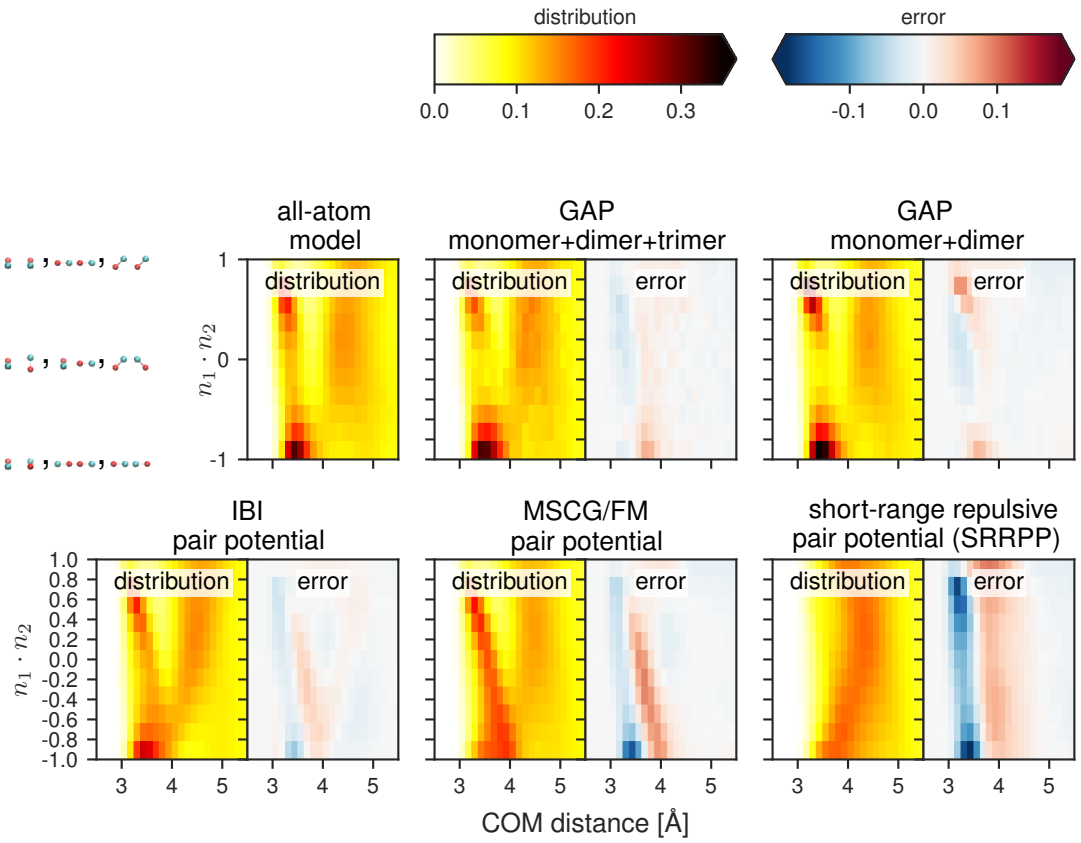

Figure 6: Orientational distribution in bulk methanol. We show the distribution of the dot product between the normalised vectors that correspond to the directions of the CG bonds, as a function of the COM distance between pairs of methanol molecules. In the top left, we show three illustrative dimer configurations each for $n_1 \cdot n_2 = +1$ (parallel), 0 (perpendicular), and -1 (antiparallel). The distributions have been normalised along the distance axis according to the RDF convention; the sum over orientations corresponds to the COM RDF. The right-hand plot of each pair shows the deviation from the all-atom result.

clearly shows a deviation from a uniform distribution in the all-atom model up to about $10\,\text{Å}$. This is reproduced by the dimer-descriptor GAP model over the whole range of COM separations and can still be seen at a COM distance of $10\,\text{Å}$, even though the (COM-based) descriptor cutoff is only $8\,\text{Å}$. In contrast, none of the pair potentials can reproduce the non-uniform orientational distribution beyond the first peak at $6\,\text{Å}$, even though the pair potential cutoff is $10\,\text{Å}$. This demonstrates that GAP based on the many-body dimer descriptor can reproduce the all-atom distribution much better than the pairwise spline-based potential. Note that in this case, the main contribution to the orientational distribution stems from excluded volume, as demonstrated by the short-range repulsive pair potential (SRRPP) having distributions very similar to the force-matching (MSCG/FM) and structure-matching (IBI) pair potentials.

4.3 Benzene dimer in water

We studied a benzene dimer solvated by 874 water molecules, in a cubic box with side length $L = 29.547\,\text{Å}$ after equilibration. The water molecules mediate an effective interaction between the benzene molecules, influenced by whether an integral number of water molecules fits in the space between the benzenes. This results in an effective hydrophobic attraction. The distribution of water molecules and the centre-of-mass mean force and PMF for the pair of benzene molecules is shown in Fig. 8. To capture the longer-range effective interactions in the water-solvated system, we extended the cutoff to $11\,\text{Å}$ for both GAP and CG pair potentials. In all other respects, the GAP hyperparameters are the same as for bulk benzene, given in Table S3, including the same SRRPP baseline.

For comparison, we also include a "traditional" implicit solvent model of the solvated benzene dimer in which the solute is described in atomistic detail, based on the Generalised Born model for solvation effects.[48] In Fig. 9 we compare the distance distribution of these CG potentials with the all-atom model. Only the many-body dimer GAP is able to capture the effective interactions and reproduce the all-atom distance distributions. Note that the water molecules in the all-atom simulation mediate an effective interaction between the benzene molecules even beyond the all-

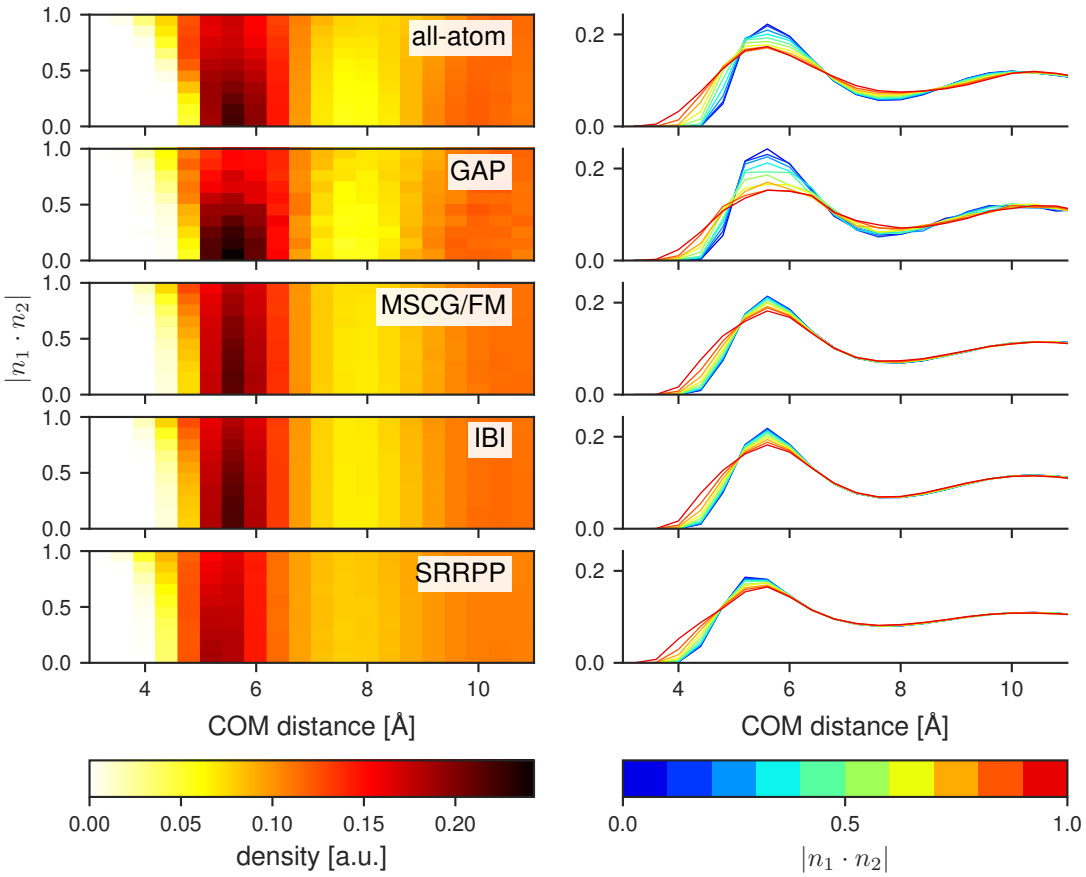

Figure 7: Orientational distribution in bulk benzene. This shows the distribution of plane orientations for pairs of benzene molecules as a function of their COM distance. The relative plane orientation is measured by the dot product $|n_1 \cdot n_2|$ between plane normal vectors; due to the symmetry of benzene, only the absolute value matters. A value of 1 corresponds to parallel orientation; a value of 0 corresponds to perpendicular orientation. This is an average over all pairs of molecules in the simulation box. The distributions have been normalised according to the RDF convention along the distance axis. *Left:* 2D distribution. *Right:* horizontal slices through this distribution, i.e. each line describes the distribution of a (small interval around a) value of $|n_1 \cdot n_2|$, as indicated by the color bar on the bottom right.

atom force field cutoff of 9 Å.

In Fig. 10 we show the orientational distribution of the plane normal vectors via their dot product, similar to the visualisation in Fig. 6. This demonstrates that the dimer-descriptor GAP model can capture the effective interactions both qualitatively and with a significantly smaller quantitative error than the other methods, whereas both CG pair potentials and the atomistic implicit solvent model fail to represent the distributions given by the all-atom model with explicit water even qualitatively.

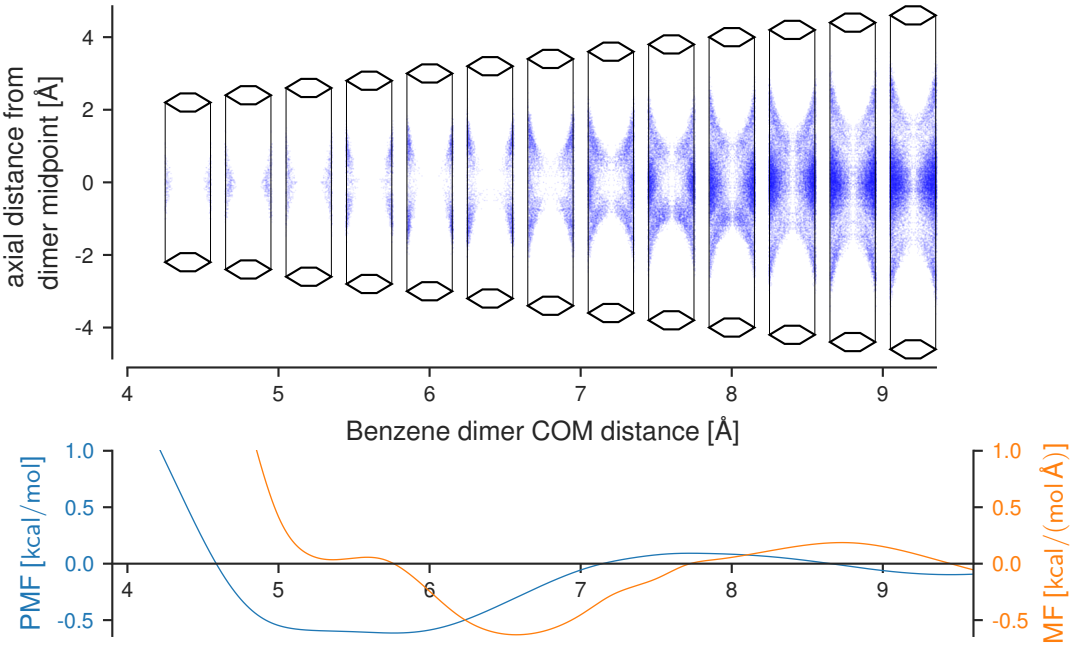

Figure 8: Water-solvated benzene dimer: distribution of water molecules. For a range of COM distances between the benzene molecules, the distribution of the water molecules is shown in blue, measured in terms of axial and radial distance of the oxygen atom from the line connecting the benzene COMs. Each box corresponds to a radial view of a cylinder between the two benzene COMs with radius 2.5 Å; the black hexagons indicate the benzene molecules and are centred according to the COM distance. Up to a COM separation of 7 Å the water molecules only encroach the edges of the cylinder between the two benzene molecules. Beyond 8 Å there is always a water molecule between them. The sharing of the hydrophobic pocket with no water between two benzene molecules effectively gives them an additional attraction. This is reflected by the behaviour of mean force (MF) and PMF for the benzene COM separation.

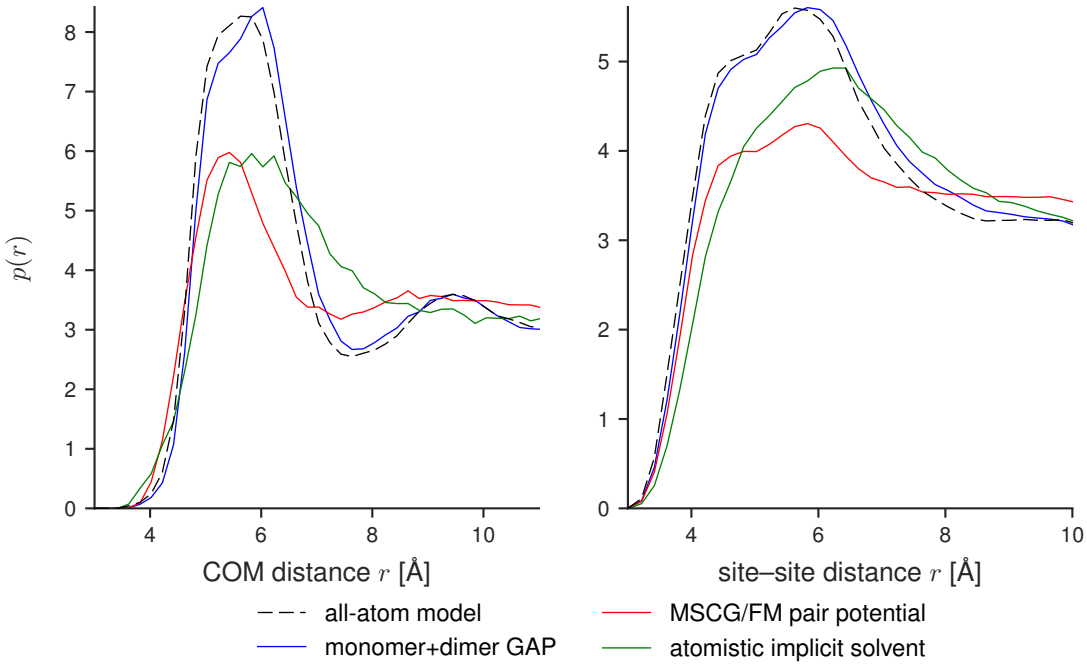

Figure 9: Water-solvated benzene dimer: distance distributions. We compare the dimer GAP, the MSCG/FM pair potential, and the atomistic implicit solvent model with the all-atom reference model that explicitly includes water. In each case, a half-harmonic restraint was applied to the COM distance starting at $12\,\text{Å}$, with force constant $50\,\text{kcal}/(\text{mol}\,\text{Å}^2)$.

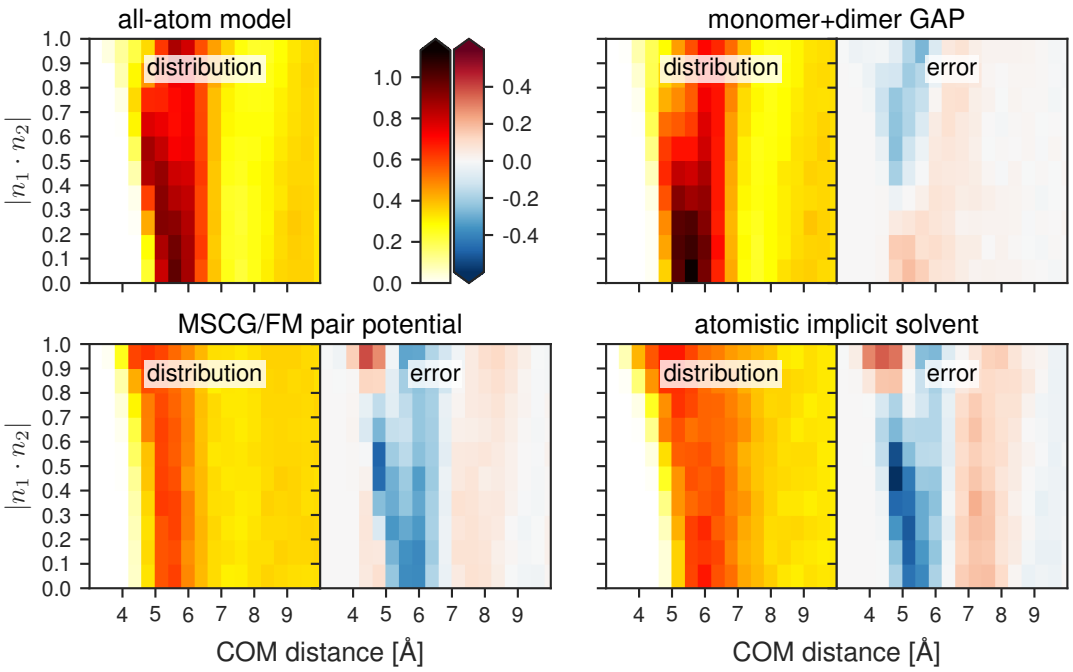

Figure 10: Water-solvated benzene dimer: orientational distribution. Plane orientation measured by the absolute value of the dot product between the plane normal vectors, $|n_1 \cdot n_2|$, as a function of the COM distance. Uniformly random orientations would lead to a uniform profile of the dot product at a fixed distance. The distributions have been normalised along the distance axis according to the RDF convention. The right-hand plot of each pair shows the deviation from the all-atom result.

5 Conclusions

We have shown that our GAP framework can recover the many-body PMF of a coarse-grained model to a high accuracy. GAP-CG is significantly better at reproducing distributions and correlations (Radial and Angular Distribution Functions as well as orientational distributions) of the underlying all-atom model than pair potentials, irrespective of how those pair potentials are generated. This shows that currently available CG models have not yet reached the limits of capability given the choice of coarse graining, and that taking multi-site effects into account beyond pairs of sites is a possible way forward. Our GAP-CG models take two somewhat independent steps in describing multi-body effects. The first is the decomposition of the total free energy of the molecular liquid into monomer, dimer and trimer terms, entirely analogously to how this is usually done when modeling the potential energy surface of condensed phase molecular systems. Secondly, in separately approximating each of these terms, we use their full dimensionality, which allows systematic convergence to the free energy contributions of monomers, dimers and trimers. It is in this second step that beyond-two-site correlations are explicitly taken into account, in our case in fact to all orders that are present in each term. Although costs would substantially increase, and corresponding benefits represent significantly diminishing returns, *in principle* this approach can be used to converge to the total free energy surface by including further n-molecule terms.

The GAP-CG approach has no inherent restrictions as to the molecules' conformational flexibility; the comparatively rigid methanol and benzene molecules are just initial examples. As the descriptors can include both the intermolecular and the intramolecular distances jointly, GAP has the flexibility to accurately capture the interactions of flexible molecules. In principle, it would be possible to describe molecules with any number of sites using GAP-CG, but this massively increases the dimensionality of the descriptors and thereby the complexity of fitting the local contributions, and hence the GAP-CG approach is likely to remain restricted to a few sites per molecule. However, using full-dimensional fits to *molecular* interactions rather than pair or three-body potentials between *sites* allows much more complicated functions, and for this reason we anticipate that it is possible to get good CG models with fewer sites that stand for much larger groups of

atoms than typically used in bottom-up CG approaches.

Compared to simple spline or tabulated pair potentials, describing multi-site correlations will always incur increased computational costs. Depending on the application, this tradeoff may or may not be worthwhile. The fundamental message of our work is that CG models can be improved, and so this tradeoff can now be made. Our current implementation within the QUIP package is a reference, driven by convenience and ease of development, rather than the desire to maximise computational efficiency. However, in conjunction with the larger time step that can be used in coarse-grained simulations, even the current GAP-CG implementation leads to simulations that are an order of magnitude faster than using an all-atom model. For a production simulation, the GAP model could be hard-coded rather than realised within QUIP, thus freeing it of the baggage that comes with the generality of QUIP that allows for arbitrary descriptors and kernels. Furthermore, there may very well be formulations of intermediate complexity and computational cost that go beyond pair potentials, but do not aim for the total flexibility of full-dimensional descriptions of the free energy terms and the Gaussian kernels.

Acknowledgement

S. T. J. acknowledges support from EPSRC/NanoDTC and Mike Payne. G. C. acknowledges support from EPSRC under grant no. EP/J010847/1.

Supporting Information Available

The Supporting Information includes the equilibration protocol used for the Molecular Dynamics simulations, GAP hyperparameter values, and plots of the pair potentials referenced in this paper.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

A Gaussian process regression for local terms

Gaussian process regression can be considered the Bayesian, probabilistic equivalent of kernel ridge regression (KRR). KRR is a common approach to function interpolation/fitting, in which an unknown function $f(x)$ is approximated by a sum of basis functions given by the *kernel* $k(x, x')$,

$$f(x) = \sum_i \alpha_i k(x_i, x), \tag{19}$$

where x_i are the basis points (which, in general, can be multi-dimensional vectors). A commonly used kernel is the squared-exponential kernel,

$$k(x, x') = \delta^2 \exp\left(- \frac{|x - x'|^2}{2\theta^2} \right). \tag{20}$$

We can determine the coefficients α_i by fitting to observations (x_j, y_j) of function values y_j at positions x_j. We require

$$y_j = \sum_i \alpha_i k(x_i, x_j). \tag{21}$$

This can be written as a matrix equation,

$$\boldsymbol{y} = K\boldsymbol{\alpha}, \tag{22}$$

where $K_{ij} = k(x_i, x_j)$, and we can solve for $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = K^{-1}\boldsymbol{y}. \tag{23}$$

However, this assumes that the observations can be exactly represented by the basis functions; in practice, K^{-1} may not exist. To overcome this problem, we need to regularise the linear system: instead of the matrix K directly, we define the matrix $C = K + \lambda I$, where I is the identity matrix and λ is the regularisation parameter, and $\boldsymbol{\alpha} = C^{-1}\boldsymbol{y}$.

We can then predict the function value y^* at a new position x^* using

$$y^* = f(x^*) = \boldsymbol{k}_*^\intercal \boldsymbol{\alpha}, \tag{24}$$

where the elements of $\boldsymbol{k}_*$ are given by $k_{*,i} = k(x_i, x^*)$.

The same result can be obtained from a probabilistic treatment in a Bayesian framework.[41,49] The kernel $k(x, x')$ is interpreted as the covariance $\mathrm{cov}(y(x), y(x')) = \delta^2 \exp(-|x - x'|^2/2\theta^2)$ of the function values at x and x'. We can then consider the observations as drawn from a multivariate Gaussian distribution, with the prior probability of the observations (x_j, y_j) given by

$$P(\boldsymbol{y}) = \mathrm{Normal}(0, C) \propto \exp\big(-\tfrac{1}{2}\boldsymbol{y}^\intercal C^{-1} \boldsymbol{y}\big), \tag{25}$$

and $C = \langle \boldsymbol{y}\boldsymbol{y}^\intercal \rangle$ is the $N \times N$ matrix that describes the covariances between the N different observations. Though here we assume the *prior* mean to be zero, this does not restrict the predictions of the CG potential; the *posterior* mean function is non-zero and depends on the observations. The elements of C are given by the covariance function that is equivalent to the kernel, $C_{ij} = \mathrm{cov}(y_i, y_j) = k(x_i, x_j)$. Note that the covariance between function values only depends on the positions $\boldsymbol{x}$ of the observations, and not on the function values $\boldsymbol{y}$ themselves.

In practice, the observations from which we train our model are noisy. There are several contributions to this noise: there may be intrinsic noise, such as in samples of instantaneous collective forces, or calculations of mean forces in constrained or restrained MD may not be fully converged; the sum of local contributions based on a given choice of descriptors can only describe a certain subspace of the full $3N$-dimensional configuration space, and hence may not be able to exactly represent the true free energy surface, and moreover we typically include only pairs and triplets within a certain cutoff radius.

In the probabilistic framework, we can describe noisy observations $\boldsymbol{y}$ by considering them to be drawn from a probability distribution depending on the noise-free values $\boldsymbol{f}$ given by the underlying model. In the simplest case, this will be a Gaussian distribution centred on the noise-free values

with a certain variance σ_f^2 that describes the noise level:

$$P(\boldsymbol{y}|\boldsymbol{f}) = \mathrm{Normal}(\boldsymbol{f}, \sigma_f^2 I). \tag{26}$$

In the framework of Gaussian process regression, this leads to an additional noise term in the covariance between observations at training points. We assume that there is no correlation between the variances in different observations, and hence we only need to add a diagonal matrix $\sigma_f^2 I$ to our noise-free covariance matrix,

$$\mathrm{cov}(y_i, y_j) = C_{ij} + \sigma_f^2 \delta_{ij}, \tag{27}$$

where δ_{ij} is the Kronecker delta for the indices of observations. [Note that in contrast to the noise-free covariance between function values, this only depends on the *indices*, not on the positions of the observations.]

The statistical error in the observations can be measured in the training data by binning in descriptor space and calculating the variance of the observations in each bin. Note that this does not account for the *systematic* error due to lack of representation by descriptors and, to prevent overfitting, the assumed noise level should be chosen larger than the measured variance.

The joint probability of the observations (x_j, y_j) and finding a function value y^* at a test point x^* is

$$P(y^*, \boldsymbol{y}) \propto \exp\big(-\tfrac{1}{2}[\boldsymbol{y}^\intercal y^*] C_{N+1}^{-1} \begin{bmatrix} \boldsymbol{y} \\ y^* \end{bmatrix}\big), \tag{28}$$

in which C_{N+1} is the extended covariance matrix of observations that also includes the test point,

$$C_{N+1} = \begin{bmatrix} [C] & [\boldsymbol{k}_*] \\ [\boldsymbol{k}_*^\intercal] & [\kappa_{**}] \end{bmatrix}, \tag{29}$$

where $\kappa_{**} = \mathrm{cov}(y^*, y^*) = k(x^*, x^*)$ is the self-covariance and $\boldsymbol{k}_* \equiv \langle y^* \boldsymbol{y} \rangle$ is the vector of

covariances between observations and the function value that we want to predict for the test point. The conditional probability for finding y^* at x^*, given the observations (x_j, y_j), is again a Gaussian distribution, with

$$P(y^*|\boldsymbol{y}) \propto \exp\left(-\frac{|y^* - \hat{f}_*|^2}{2\hat{\sigma}_*^2}\right). \tag{30}$$

The centre of this posterior distribution can be derived analytically,[41,49] resulting in

$$\hat{f}_* = \boldsymbol{k}_*^\intercal C^{-1} \boldsymbol{y} \tag{31}$$

as before. The key difference between kernel ridge regression and Gaussian process regression is the interpretation of regularisation. In KRR, there is an empirical regularisation parameter λ that needs to be optimised. In the Bayesian framework of Gaussian process regression, the regularisation parameter can be identified with the standard deviation of noisy observations, $\lambda = \sigma_f^2$; i.e. it is a *physical* quantity that is an inherent property of the data. Moreover, in Gaussian process regression we can determine the uncertainty in the prediction, $\hat{\sigma}_*$:

$$\hat{\sigma}_*^2 = \kappa_{**} - \boldsymbol{k}_*^\intercal C^{-1} \boldsymbol{k}_*. \tag{32}$$

Note that the underlying noise-free model and the noisy predicted observation y^* have the same mean. The variance of the latter is larger by the variance of the noise model, $\mathrm{Var}[y] = \mathrm{Var}[f] + \sigma_f^2$.

The choice of kernel function determines the properties of the functions described by the prior distribution. In the following we use the squared-exponential kernel (20), which leads to smooth, infinitely differentiable functions with length scale θ and function value scale δ, appropriate for interaction potentials.

A.1 Derivatives, sums, and permutations

We can straightforwardly predict derivatives of the inferred function value by taking the derivative of Eq. (31) with respect to components x^*_m of the test point position:

$$\frac{\partial \hat{f}_*}{\partial x^*_m} = \frac{\partial \boldsymbol{k}_*^\intercal}{\partial x^*_m} C_N^{-1} \boldsymbol{y}, \tag{33}$$

where the partial derivative on the right-hand side is the derivative of the covariance function. For the squared-exponential kernel (20), this is given by

$$\frac{\partial k_{*,j}}{\partial x^*_m} = \frac{x_{j,m} - x^*_m}{\theta^2_m} k(\boldsymbol{x}_j, \boldsymbol{x}^*). \tag{34}$$

In KRR, we would learn from derivatives by simply taking the derivative of Eq. (19); this would lead to

$$\nabla_x f = \sum_i \alpha_i \nabla_x k(x, x_i).$$

However, the basis functions defined by the kernel are maximal at their centre and hence the derivative is zero, so we would attempt to fit the derivatives using functions that are zero at the observations. The Bayesian framework shows us what the correct basis functions should be, which has been discussed in the literature.[22–26,50,51]

Because Gaussian process regression is based on the covariances between observations, it can easily be extended to more complex observations than just function values, for example sums of function values and derivatives.[50] We want to describe the free energy as a sum of local terms that are part of our model but cannot be observed directly. Being able to train from sum observations, we can *infer* the "hidden" local contributions to the free energy due to monomer, dimer, and trimer terms. Moreover, we can only observe the derivatives of the total free energy, the mean forces.

We model the free energy as follows:

$$U_{\text{CG}} = \delta_1 \sum W_{\text{mono}}(\boldsymbol{D}_{\text{mono}}) + \delta_2 \sum W_{\text{dimer}}(\boldsymbol{D}_{\text{dimer}}) + \delta_3 \sum W_{\text{trimer}}(\boldsymbol{D}_{\text{trimer}}) + \varepsilon, \tag{35}$$

where the strengths of the different contributions to the CG potential are explicitly given by the δ prefactors and the W functions all have unit scale, and $\varepsilon \sim \mathrm{Normal}(0, \sigma_f^2)$ is a random variable that describes the noise/error. Note that this error includes both any noise due to the measurement of the gradients of the free energy, ∇A, as well as the deviation between A and what our model U_{CG} is able to describe.

Assuming for a moment just a single type of descriptors (e.g., dimers) then the covariance between free energy values U_{CG}^N and U_{CG}^M for two configurations N and M is given by

$$\begin{aligned}\mathrm{cov}(U_{\mathrm{CG}}^N, U_{\mathrm{CG}}^M) = \langle U_{\mathrm{CG}}^N U_{\mathrm{CG}}^M \rangle &= \left\langle \big(\delta \sum_{i\in N} W(\boldsymbol{D}_i) + \varepsilon\big)\big(\delta \sum_{j\in M} W(\boldsymbol{D}_j) + \varepsilon\big) \right\rangle \\ &= \delta^2 \sum_{i\in N}\sum_{j\in M} \langle W(\boldsymbol{D}_i) W(\boldsymbol{D}_j)\rangle + \langle \varepsilon^2 \rangle = \sum_{i\in N}\sum_{j\in M} k(\boldsymbol{D}_i, \boldsymbol{D}_j) + \sigma_f^2, \quad (36)\end{aligned}$$

where i and j iterate over the descriptors found in the configurations N and M, respectively (e.g., all pairs of molecules within the descriptor's cutoff), and the "signal variance" δ^2 is included as part of the kernel k, cf. Eq. (20). To infer the free energy from forces, we need the covariance between two derivative observations, which is given by

$$\mathrm{cov}\left(\frac{\partial U_{\mathrm{CG}}^N}{\partial R_k}, \frac{\partial U_{\mathrm{CG}}^M}{\partial R_l}\right) = \frac{\partial^2 \langle U_{\mathrm{CG}}^N U_{\mathrm{CG}}^M \rangle}{\partial R_k \partial R_l} = \sum_{i\in N}\sum_{j\in M}\sum_{\alpha,\beta} \frac{\partial D_{i\alpha}}{\partial R_k} \frac{\partial^2 k(\boldsymbol{D}_i, \boldsymbol{D}_j)}{\partial D_{i\alpha} \partial D_{j\beta}} \frac{\partial D_{j\beta}}{\partial R_l}, \quad (37)$$

where k and l index CG coordinates in configurations N and M, and α and β iterate over the dimensions of the descriptors $\boldsymbol{D}_i$ and $\boldsymbol{D}_j$, respectively. From these covariances we build up the covariance matrix C'', which has a block-diagonal structure, with separate blocks for each configuration.

When we want to include several different types of descriptors in our model, as will generally be the case, we would in principle have to include the covariance between different types of descriptors, e.g., $\langle W_{\mathrm{mono}}(\boldsymbol{D}_{\mathrm{mono}}) W_{\mathrm{dimer}}(\boldsymbol{D}_{\mathrm{dimer}})\rangle$. While such a covariance is likely to exist, we cannot easily describe or determine this covariance a-priori. Hence, we make the approximation that the covariances between different types of descriptors are zero. This means that the total covariance

will simply be a sum of the covariances due to different descriptor types.

Crucially, descriptors of higher complexity may incorporate the flexibility of lower-dimensional descriptors. For example, the dimer descriptor also contains intramolecular distances, which would allow it to effectively describe the monomer term. However, only using the dimer descriptor would lead to a more complex contribution to the free energy surface, which makes it harder to infer from the training data in such high-dimensional spaces. Instead, we would like the dimer descriptor to only describe what cannot be described by the monomer term, and in fact we expect the monomer term to have the largest contribution to the overall free energy surface. The interaction free energy described by the dimer descriptor should not include the "self-energy" of the monomer term. In our framework, each descriptor has its own variance, which describes the amplitude of a descriptor's contribution to the overall free energy surface. In the squared-exponential kernel this is set by the δ prefactors. By setting the different variance hyperparameters we can define the (expected, prior) relative importance of the different terms, thus enabling our framework to separate the different contributions.

The training data, i.e. observations of the mean force directly or of the stochastic mean force estimator given by the ICF, are collected in the vector $\boldsymbol{y}'$. We can then determine the coefficients

$$\boldsymbol{\alpha} = [C'']^{-1}\boldsymbol{y}',$$

and the overall prediction of a local contribution to the free energy is given by

$$W(\boldsymbol{D}) = f(\boldsymbol{x}) = \boldsymbol{k}_*^\intercal \boldsymbol{\alpha}$$

as before.

When the descriptor is not already invariant with respect to permutations of equivalent CG sites, as is the case for our monomer-based descriptors, we can give permutational symmetry to the model by summing the covariance over all valid permutations $\hat{P}$ of the components of the

descriptor $\boldsymbol{D}$ that are part of the permutation group of the molecule/cluster, $\mathcal{P}$:

$$\tilde{C}(\boldsymbol{D}, \boldsymbol{D}') = \sum_{\hat{P} \in \mathcal{P}} C(\boldsymbol{D}, \hat{P}\boldsymbol{D}'), \tag{38}$$

which then needs to be renormalised so that $C'(\boldsymbol{D}, \boldsymbol{D}) = 1$:

$$C'(\boldsymbol{D}, \boldsymbol{D}') = \frac{\tilde{C}(\boldsymbol{D}, \boldsymbol{D}')}{\sqrt{\tilde{C}(\boldsymbol{D}, \boldsymbol{D})}\sqrt{\tilde{C}(\boldsymbol{D}', \boldsymbol{D}')}}. \tag{39}$$

Note that $\hat{P}\boldsymbol{D}'$ describes the distances resulting from permuted (equivalent) CG sites, not permutations of the distances directly.

A.2 Sparsification

The Gaussian process regression as introduced so far relies on the full covariance matrix between observations. For large numbers of observations, especially given that each configuration involves $3N$ derivative observations, this can become very expensive and, moreover, is unnecessary. In practice, the descriptors found in the neighbourhood environments are highly correlated, which motivates the use of *sparse* Gaussian process regression. For a more in-depth discussion of sparsification we refer to the paper by Quiñonero-Candela and Rasmussen.[52] Sparsification can be considered as a projection onto a set of "sparse points" or "inducing points", representative monomer, dimer, and trimer configurations, whose total number can be much smaller than the total number of observations. The predictive posterior probability is then approximated by

$$P(f_*|\boldsymbol{y}) \approx P(f_*|\tilde{\boldsymbol{y}})P(\tilde{\boldsymbol{y}}|\boldsymbol{y}), \tag{40}$$

where $\tilde{\boldsymbol{y}}$ are the function values at the sparse points $\tilde{\boldsymbol{x}}$, which we do not observe, but we can condition on them, analogously to the standard literature. Even though we can only observe derivatives of the total free energy, we can base the sparsification on the function values of the local terms at

the sparse points, as they are inferred from the actual observations.

To represent the sparse Gaussian process regression, we need the matrix of covariances between sparse points,

$$(C_{SS})_{ss'} = \langle W_s W_{s'} \rangle, \tag{41}$$

and the matrix of covariances between sparse points and training observations,

$$(C_{ST})_{s\tau} = \langle W_s \frac{\partial A_t}{\partial R_k} \rangle, \tag{42}$$

where $\tau = (t, k)$ is the index of the mean force k in configuration t, and $C_{TS} = C_{ST}^{\intercal}$.

In our work we use the Deterministic Training Conditional approximation,[53] a standard technique to reduce the burden of solving the linear system. In this approximation, the predicted function value is given by

$$\hat{f}_* = \boldsymbol{k}_*^{\intercal} Q^{-1} \tilde{\boldsymbol{y}} \tag{43}$$

where

$$Q = C_{SS} + C_{ST} \Lambda_{TT}^{-1} C_{TS} \tag{44}$$

replaces the covariance matrix between observations, and

$$\tilde{\boldsymbol{y}} = C_{ST} \Lambda_{TT}^{-1} \boldsymbol{y} \tag{45}$$

describes the function values at the sparse points predicted by the training data. $\Lambda_{TT} = \sigma_f^2 I$ is a diagonal matrix with the noise variance of the mean force observations on the diagonal. The resulting prediction can again be written as a dot product, similar to Eq. (31), $\hat{f}_* = \sum_s \tilde{\alpha}_s k(x, x_s)$, where $\tilde{\boldsymbol{\alpha}} = Q^{-1} \tilde{\boldsymbol{y}}$ are the sparse coefficients that can be precomputed in the training phase.

A.3 Inhomogeneous length scales / transfer function

Gaussian process regression commonly assumes a stationary covariance, that is, the covariance only depends on the distance $|x - x'|$ between two points and not their absolute positions. This is the case for the popular squared-exponential kernel (20), which has a constant length scale θ. However, the interactions between atoms and, by extension, between CG sites do not have a single length scale. At long range/large intermolecular separations, atomic as well as CG interactions tend to be weak and slowly changing. At short range, overlap of atomic electron clouds leads to a strong, quickly changing repulsion.

If the constant length scale is appropriate at short range, then we will overfit to noise at long range. If the constant length scale is appropriate at long range, then at short range we will induce correlations between force observations that are actually very different from each other, leading to oscillations or "ringing" in energies and forces.

While it would be possible to introduce a non-stationary covariance[54] that depends on the distance, a simpler solution is to simply stretch the coordinates at short range, leading to an effective length scale that is similar to that at long range. We describe this by writing the local free energy contribution as

$$W(x) = \tilde{W}\big(T(x)\big), \tag{46}$$

where for simplicity we consider the case of a one-dimensional descriptor x. $T(x)$ is the transfer function that describes the stretched coordinate. The derivative of the local free energy contribution is then given by

$$\frac{\mathrm{d}W}{\mathrm{d}x} = \frac{\mathrm{d}\tilde{W}}{\mathrm{d}\tilde{x}} \frac{\mathrm{d}T(x)}{\mathrm{d}x}. \tag{47}$$

The derivative of the transfer function describes the local stretch factor. We take the long-range length scale as the base line: for $x \gg r_0$, the stretch factor should go to 1; for $x \ll r_0$, the derivative should go to the short-range stretch factor α between short-range and long-range length scale, with a smooth transition between the two regimes. This is achieved by the following functional form

for the derivative:

$$\frac{\mathrm{d}T}{\mathrm{d}x} = (1-\alpha)\frac{1}{2}\left[\tanh\left((x-r_0)/\gamma\right)+1\right]+\alpha, \tag{48}$$

where γ is the transition width. We can obtain the actual transfer function by integrating Eq. (48). This results in

$$T(x) = (1-\alpha)\frac{1}{2}\left[\gamma\log\left(2\cosh((x-r_0)/\gamma)\right)+x+r_0\right]+\alpha x. \tag{49}$$

This transfer function depends on three parameters: r_0 is the mid-point for the transition between the length scales, γ is the extent of the transition region either side of r_0, and α is the stretching factor.

In the present work, we apply the transfer function only to the non-bonded distances in the dimer descriptor. Bonded distances only have a single length scale, and the trimer interactions are expected to be comparatively smooth and slowly varying already.

References

(1) Shell, M. S. The Relative Entropy is Fundamental to Multiscale and Inverse Thermodynamic Problems. *J. Chem. Phys.* **2008**, *129*, 144108.

(2) Izvekov, S.; Voth, G. A. A Multiscale Coarse-Graining Method for Biomolecular Systems. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.

(3) Izvekov, S.; Voth, G. A. Multiscale Coarse Graining of Liquid-State Systems. *J. Chem. Phys.* **2005**, *123*, 134105.

(4) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The Multiscale Coarse-Graining Method. I. A Rigorous Bridge between Atomistic and Coarse-Grained Models. *J. Chem. Phys.* **2008**, *128*, 244114.

(5) Lu, L.; Izvekov, S.; Das, A.; Andersen, H. C.; Voth, G. A. Efficient, Regularized, and Scalable Algorithms for Multiscale Coarse-Graining. *Journal of Chemical Theory and Computation* **2010**, *6*, 954–965.

(6) Rudzinski, J. F.; Noid, W. G. Coarse-Graining Entropy, Forces, and Structures. *J. Chem. Phys.* **2011**, *135*, 214101.

(7) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. Multiscale Coarse-Graining and Structural Correlations: Connections to Liquid-State Theory. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.

(8) Mullinax, J. W.; Noid, W. G. Generalized Yvon–Born–Green Theory for Molecular Systems. *Phys. Rev. Lett.* **2009**, *103*, 198104.

(9) Mullinax, J. W.; Noid, W. G. A Generalized-Yvon–Born–Green Theory for Determining Coarse-Grained Interaction Potentials. *J. Phys. Chem. C* **2010**, *114*, 5661–5674.

(10) Rudzinski, J. F.; Noid, W. G. Investigation of Coarse-Grained Mappings via an Iterative Generalized Yvon–Born–Green Method. *J. Phys. Chem. B* **2014**, *118*, 8295–8312.

(11) Pagonabarraga, I.; Frenkel, D. Dissipative Particle Dynamics for Interacting Systems. *J. Chem. Phys.* **2001**, *115*, 5015–5026.

(12) Moore, J. D.; Barnes, B. C.; Izvekov, S.; Lísal, M.; Sellers, M. S.; Taylor, D. E.; Brennan, J. K. A Coarse-Grain Force Field for RDX: Density Dependent and Energy Conserving. *J. Chem. Phys.* **2016**, *144*, 104501.

(13) Sanyal, T.; Shell, M. S. Coarse-Grained Models Using Local-Density Potentials Optimized with the Relative Entropy: Application to Implicit Solvation. *J. Chem. Phys.* **2016**, *145*, 034109.

(14) Dama, J. F.; Jin, J.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 3. Coarse-Grained Sites with Rapid Local Equilibrium of Internal States. *J. Chem. Theory Comput.* **2017**, *13*, 1010–1022.

(15) Molinero, V.; Moore, E. B. Water Modeled as an Intermediate Element between Carbon and Silicon. *J. Phys. Chem. B* **2009**, *113*, 4008–4016.

(16) Larini, L.; Lu, L.; Voth, G. A. The Multiscale Coarse-Graining Method. VI. Implementation of Three-Body Coarse-Grained Potentials. *J. Chem. Phys.* **2010**, *132*, 164107.

(17) Das, A.; Andersen, H. C. The Multiscale Coarse-Graining Method. IX. A General Method for Construction of Three Body Coarse-Grained Force Fields. *J. Chem. Phys.* **2012**, *136*, 194114.

(18) Liu, Y.; Ichiye, T. Soft Sticky Dipole Potential for Liquid Water: A New Model. *J. Phys. Chem.* **1996**, *100*, 2723–2730.

(19) Dijkstra, M.; van Roij, R.; Evans, R. Phase Diagram of Highly Asymmetric Binary Hard-Sphere Mixtures. *Phys. Rev. E* **1999**, *59*, 5744–5771.

(20) Bolhuis, P. G.; Louis, A. A.; Hansen, J. P. Many-Body Interactions and Correlations in Coarse-Grained Descriptions of Polymer Solutions. *Phys. Rev. E* **2001**, *64*, 021801.

(21) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-Learning Approach for One- and Two-Body Corrections to Density Functional Theory: Applications to Molecular and Condensed Water. *Phys. Rev. B* **2013**, *88*, 054104.

(22) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(23) Szlachta, W. J.; Bartók, A. P.; Csányi, G. Accuracy and Transferability of Gaussian Approximation Potential Models for Tungsten. *Phys. Rev. B* **2014**, *90*, 104108.

(24) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids Across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

(25) Deringer, V. L.; Csányi, G.; Proserpio, D. M. Extracting Crystal Chemistry from Amorphous Carbon Structures. *ChemPhysChem* **2017**, *18*, 873–877.

(26) Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B* **2017**, *95*, 094203.

(27) Mones, L.; Bernstein, N.; Csányi, G. Exploration, Sampling, and Reconstruction of Free Energy Surfaces with Gaussian Process Regression. *J. Chem. Theory Comput.* **2016**, *12*, 5100–5110.

(28) John, S. T. Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials. Ph.D. thesis, University of Cambridge, 2016; `https://arxiv.org/abs/1611.09123` (accessed November 2, 2017).

(29) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. Constrained Reaction Coordinate Dynamics for the Simulation of Rare Events. *Chem. Phys. Lett.* **1989**, *156*, 472–477.

(30) Sprik, M.; Ciccotti, G. Free Energy from Constrained Molecular Dynamics. *J. Chem. Phys.* **1998**, *109*, 7737–7744.

(31) Kästner, J. Umbrella Sampling. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 932–942.

(32) Kästner, J.; Thiel, W. Bridging the Gap between Thermodynamic Integration and Umbrella Sampling Provides a Novel Analysis Method: 'Umbrella Integration'. *J. Chem. Phys.* **2005**, *123*, 144104.

(33) Lelièvre, T.; Rousset, M.; Stoltz, G. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press: London, U.K., 2010.

(34) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving Effective Mesoscale Potentials from Atomistic Simulations. *J. Comput. Chem.* **2003**, *24*, 1624–1636.

(35) Lyubartsev, A. P.; Laaksonen, A. Calculation of Effective Interaction Potentials from Radial Distribution Functions: A Reverse Monte Carlo Approach. *Phys. Rev. E* **1995**, *52*, 3730–3737.

(36) Soper, A. K. Empirical Potential Monte Carlo Simulation of Fluid Structure. *Chem. Phys.* **1996**, *202*, 295–306.

(37) Lyubartsev, A. P.; Karttunen, M.; Vattulainen, I.; Laaksonen, A. On Coarse-Graining by the Inverse Monte Carlo Method: Dissipative Particle Dynamics Simulations Made to a Precise Tool in Soft Matter Modeling. *Soft Mater.* **2002**, *1*, 121–137.

(38) Das, A.; Andersen, H. C. The Multiscale Coarse-Graining Method. VIII. Multiresolution Hierarchical Basis Functions and Basis Function Selection in the Construction of Coarse-Grained Force Fields. *J. Chem. Phys.* **2012**, *136*, 194113.

(39) Liu, P.; Shi, Q.; Daumé, H.; Voth, G. A. A Bayesian Statistics Approach to Multiscale Coarse Graining. *J. Chem. Phys.* **2008**, *129*, 214114.

(40) Mullinax, J. W.; Noid, W. G. Reference State for the Generalized Yvon–Born–Green Theory: Application for Coarse-Grained Model of Hydrophobic Hydration. *J. Chem. Phys.* **2010**, *133*, 124107.

(41) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, 2006.

(42) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. Versatile Object-Oriented Toolkit for Coarse-Graining Applications. *J. Chem. Theory Comput.* **2009**, *5*, 3211–3223.

(43) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. The Multiscale Coarse-Graining Method. II. Numerical Implementation for Coarse-Grained Molecular Models. *J. Chem. Phys.* **2008**, *128*, 244115.

(44) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L. I.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M. et al. *AMBER 12*. University of California, San Francisco, 2012.

(45) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(46) Kulhánek, P.; Mones, L.; Střelcová, Z.; Simon, I.; Fuxreiter, M.; Koča, J. PMFLib — A Toolkit for Free Energy Calculations. 8th Discussions in Structural Molecular Biology, Nové Hrady, Czech Republic. 2011.

(47) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. Simulation of Polymer Melts. I. Coarse-Graining Procedure for Polycarbonates. *Acta Polym.* **1998**, *49*, 61–74.

(48) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(49) MacKay, D. J. C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, U.K., 2003.

(50) Bartók, A. P.; Csányi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

(51) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, e1603015.

(52) Quiñonero-Candela, J.; Rasmussen, C. E. A Unifying View of Sparse Approximate Gaussian Process Regression. *Journal of Machine Learning Research* **2005**, *6*, 1939–1959.

(53) Seeger, M.; Williams, C.; Lawrence, N. Fast Forward Selection to Speed Up Sparse Gaussian Process Regression. Artificial Intelligence and Statistics 9. 2003.

(54) Gibbs, M. Bayesian Gaussian Processes for Classification and Regression. Ph.D. thesis, University of Cambridge, 1997.

Graphical TOC Entry